What uncertainty do we get?

Zhenwen Dai

11 October 2019

Probabilistic Models

- Many probabilistic models have been discussed.
- We are interested in probabilistic models because it provides how uncertain it is about its prediction.
- Uncertainty has been categorized into various names such as epistemic uncertainty, aleatoric uncertainty, model uncertainty, noise.
- What do people mean by these types of uncertainty?

Uncertainty in Discriminative Model

Regression as an example:

$$y = f(x) + \epsilon$$

A simple example, Bayesian linear regression (BLR):

$$y_i = \mathbf{w}^\top \Phi(x_i) + \epsilon_i$$

Two random variables:

$$\mathbf{w} \sim \mathcal{N}(0, \mathbb{I}), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Uncertainty in Discriminative Model

- By uncertainty, we usually mean how wide is the probabilistic distribution of the predicted variable.
- For BLR, it refers to $\text{var}(y_*) = \mathbb{E}_{p(y_*|x_*)}[(y_* - \bar{y}_*)^2]$.

If we obtain maximum likelihood estimate (MLE) of w, $\hat{w}$, the predictive distribution is

$$p(y_*|x_*, \hat{\mathbf{w}}) = \hat{\mathbf{w}}^\top \Phi(x_*) + \epsilon_*.$$

If we do Bayesian inference over w, the predictive distribution is

$$p(y_*|x_*) = \int p(y_*|x_*, \mathbf{w}) p(w|\mathbf{x}, \mathbf{y}) d\mathbf{w}.$$

Epistemic and Aleatoric Uncertainty

- Aleatoric uncertainty

Aleatoric uncertainty is also known as statistical uncertainty, and is representative of unknowns that differ each time we run the same experiment.

- Epistemic uncertainty

Epistemic uncertainty is also known as systematic uncertainty, and is due to things one could in principle know but doesn't in practice. This may be because a measurement is not accurate, because the model neglects certain effects, or because particular data has been deliberately hidden.

Epistemic and Aleatoric Uncertainty in BLR

Use BLR as an example:

$$y_i = \mathbf{w}^\top \Phi(x_i) + \epsilon_i, \quad \mathbf{w} \sim \mathcal{N}(0, \mathbb{I}), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In the usual modeling scenario,

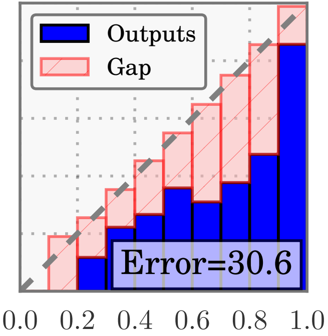
- ϵ corresponds to aleatoric uncertainty. Measured as $\text{var}(y_*) = \mathbb{E}_{p(y_*|x_*,\hat{\mathbf{w}})}[(y_* - \bar{y}_*)^2] = \sigma^2$.
- $\mathbf{w}$ corresponds to epistemic uncertainty. Measured as $\text{var}(f_*) = \mathbb{E}_{p(f_*|x_*)}[(f_* - \bar{f}_*)^2]$, where $f_* = \mathbf{w}^\top \Phi(x_*)$.

Separation of Uncertainty

- With a probabilistic model, what we care is the predictive distribution $p(y_*|x_*)$.
- The separation of epistemic and aleatoric uncertainty seems a bit artificial. Do we really need to separate them?

Probability Calibration

- It is a common question in practice whether we should trust the predictive probability.
- What does it mean when a weather forecasting method predict 70% of probability of raining.
- It is an well understood question in frequentist statistics.



Probability Calibration for Aleatoric and Epistemic Uncertainty

- Make sense for aleatoric uncertainty. It is i.i.d., $\epsilon_1, \ldots, \epsilon_N \sim p(\epsilon)$.
- Probability calibration for epistemic uncertainty?
- Does the uncertainty from the exact Bayesian posterior warrant calibrated probability on output?
- How about the measure only happened once? How shall we give a prior distribution? Would uncertainty be calibrated in this case?

Uncertainty in Decision Making

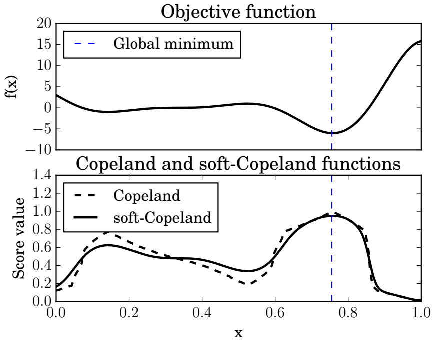
- Alternatively we may assess the quality of uncertainty by the performance of downstream tasks.
- Which uncertainty shall we use in Bayesian optimization, experimental design?

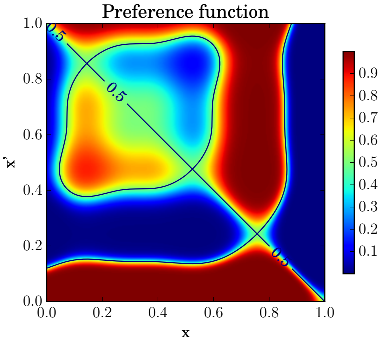
Preferential Bayesian Optimization

- Many functions that we are interested in optimizing is hard to measure:
  - user experience, e.g, UI design
  - movie/music rating
- Human are much better at comparing two things, e.g., is this coffee better than the previous one?
- To search for the most preferred option via only pair-wise comparisons.

Preference Function

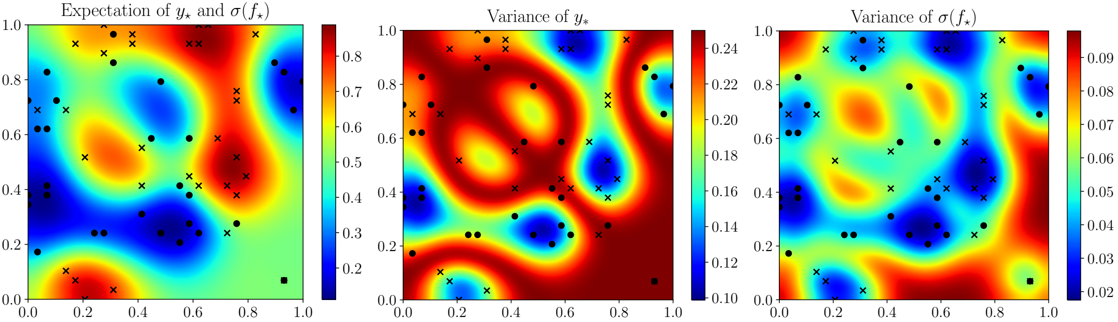
- Preference function: $p(y = 1|x, x') = \pi(x, x') = \sigma(g(x') - g(x))$.
- Copeland function: $S(x) = \frac{1}{\mathrm{Vol}(\mathcal{X})} \int_{\mathcal{X}} \mathbb{I}_{\pi(x,x') \geq 0.5} \mathrm{d}x'$.
- The minimal of a Copeland function corresponds to the most preferred choice.





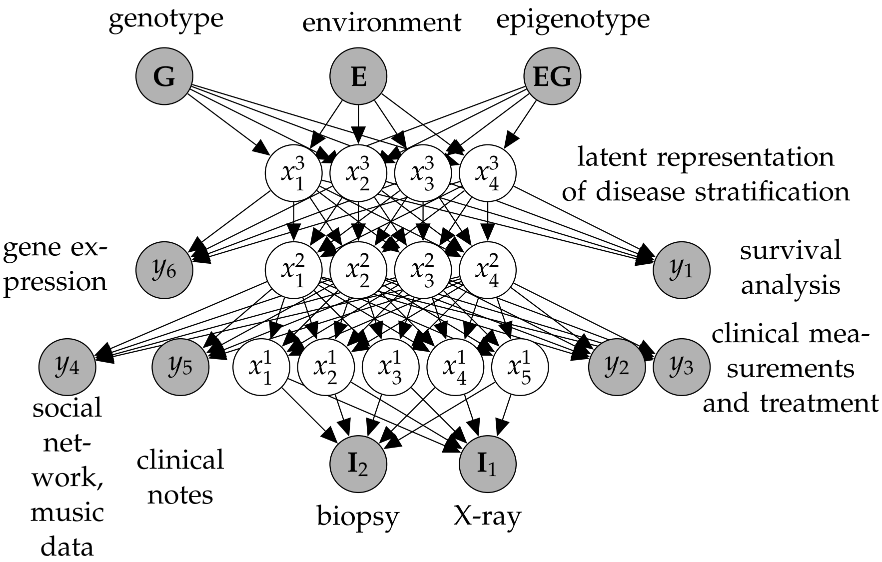
Exploration

- $p(y|x, x') = \pi(x, x')^y (1 - \pi(x, x'))^{1-y}, \quad \pi(x, x') = \sigma(f(x, x'))$.
- $\mathbb{E}[y] = \pi(x, x')$, var$(y) = \pi(x, x')(1 - \pi(x, x'))$



- Epistemic and aleatoric uncertainty are different.
- Exploration should done only with epistemic uncertainty.

What about composite model?



Disclaimer

I don't know how to categorize the uncertainty from a probabilistic generative model for unsupervised learning such as VAE, GPLVM.

Separation of Uncertainty in Complex model

- We need a systematic approach to separate epistemic and aleatoric uncertainty.
- Let's still focus on discriminative models

$$y_i = f(x_i) + \epsilon_i$$

Look back at BLR

$$y_i = \mathbf{w}^\top \Phi(x_i) + \epsilon_i, \quad \mathbf{w} \sim \mathcal{N}(0, \mathbb{I}), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Aleatoric uncertainty:

- Unknowns that differ each time we run the same experiment.

Epistemic uncertainty:

- Things one could in principle know but doesn't in practice.

One way to classify

Aleatoric uncertainty

- Unknowns that differ each time we run the same experiment.
- Independence among data points

$$y_i = (x_i, h_i)$$

Epistemic uncertainty

- Things one could in principle know but doesn't in practice.
- Global variable

$$y_i = (x_i, h)$$

Variables Shared by a Subset of Data Points

Aleatoric uncertainty

$$y_i = (x_i, h_i)$$

Epistemic uncertainty

$$y_i = (x_i, h)$$

What about something in between?

$$y_i = (x_i, h_{z(i)}), \quad z : \{1, \dots, N\} \rightarrow \{1, \dots, C\}$$

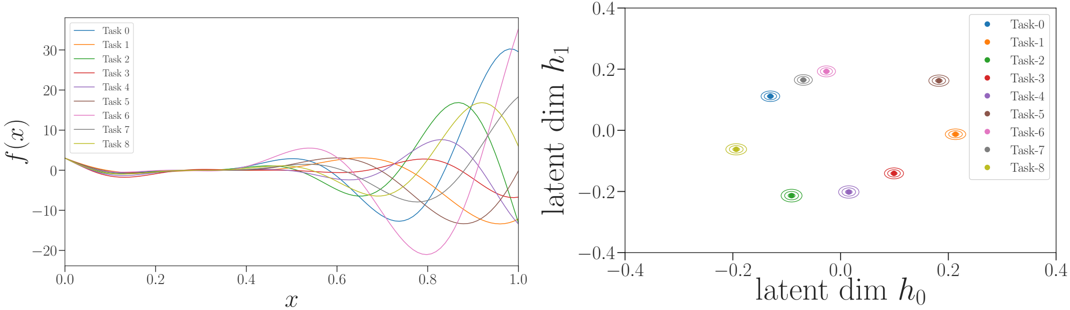
An example: Multi-output GP

Also known as Intrinsic Coregionalization

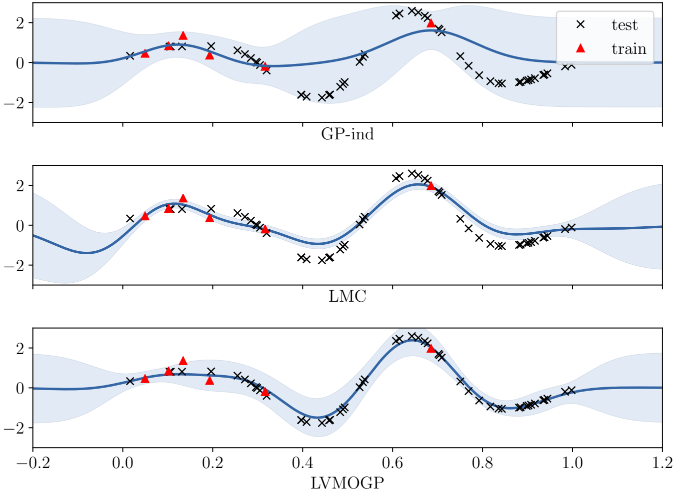
- Each input location corresponds to C different output dimensions.
- $\mathbf{f} = (f_{11}, \ldots, f_{1N}, \ldots, f_{C1}, \ldots, f_{CN})^\top$.
- $\mathbf{f}|\mathbf{X} \sim \mathcal{N}(0, \mathbf{B} \otimes \mathbf{K}), \quad \mathbf{B} \in \mathbb{R}^{C \times C}, \quad \mathbf{K} \in \mathbb{R}^{N \times N}$.

Latent variable multi-output GP

- Assume $\mathbf{B}$ is a covariance matrix computed according to a kernel function $k(\cdot,\cdot)$ over a set of variable $\mathbf{h}_1,\ldots,\mathbf{h}_C$.
- $\mathbf{h}_i$ is a latent variable, $\mathbf{h}_i \sim \mathcal{N}(0,\mathbf{I})$.



Latent variable multi-output GP



Epistemic or aleatoric?

- For multi-task learning, one output correspond to a task. The uncertainty associated with $\mathbf{h}_i$ is epistemic uncertainty of the task.
- What if only one observation can be collected for each task? It becomes aleatoric!
- A better way to see it may be epistemic within the group and aleatoric for other groups.

Soft group assignment

Let's see a more confusing case by softening the group assignment.

- The covariance of data points within a group is a bias kernel

$$\mathbf{B}_{11} = b_{11} \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}.$$

- Augment the model with one $\mathbf{h}_i$ for each data point $\mathbf{x}_i$,

$$y_i = f(\mathbf{x}_i, \mathbf{h}_i),$$

  the covariance matrix is $\mathbf{B} \odot \mathbf{K}$. The joint distribution $p(\mathbf{h}_1, \ldots, \mathbf{h}_N)$ correlates.
- A trivial case would be the degenerate distribution $\mathbf{h}_1 = \ldots = \mathbf{h}_N = \epsilon, \epsilon \sim p(\epsilon)$.

Continuous Learning

An example of previous model is a model for continuous learning.

- Data points arrives with different time, $\mathbf{x}_1, \dots, \mathbf{x}_T$ and $\mathbf{y}_1, \dots, \mathbf{y}_T$.
- The underlying function may change over time $f_1(\cdot), \dots, f_T(\cdot)$.
- We can construct such a model in the above form by constructing a state-space model,

$$p(\mathbf{h}_1, \dots, \mathbf{h}_T) = p(\mathbf{h}_1) \prod_{t=2}^{T} p(\mathbf{h}_t | \mathbf{h}_{t-1})$$

- Are $\mathbf{h}_1, \dots \mathbf{h}_T$ epistemic or aleatoric?

Summary

- Epistemic and aleatoric uncertainty and their role in decision making.
- "outliner" models that are hard to be classified.

Thoughts:

- Looking at the uncertainty of the output variable may be the best way.